DOCUMENT RESUME

ED 176 589 FL 010 823

AUTHOR	Jarvis, Gilbert A.; Adams, Shirley J.
TITLE	Evaluating a Second Language Program. Language in Education: Theory and Practice, No. 19.
INSTITUTION	ERIC Clearinghouse on Languages and Linguistics, Arlington, Va.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Sep 79
NOTE	44p.
AVAILABLE FROM	Center for Applied Linguistics, 1611 N. Kent Street, Arlington, Virginia 22209 ($2.95)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Classroom Observation Techniques; Data Analysis; Data Collection; *Educational Objectives; *Evaluation Methods; Guidelines; Instructional Materials; Language Instruction; *Language Programs; Language Tests; Measurement Instruments; *Program Evaluation; *Second Language Learning; Student Behavior; *Student Evaluation; Teacher Behavior
IDENTIFIERS	Information Analysis Products

ABSTRACT

This introduction to second language program evaluation begins with a rationale outlining the major issues involved in program evaluation. Procedures for evaluation are described, and examples are offered of goal statements, sub-goal statements, curriculum objectives, and learning objectives. Guidelines for judging a statement of aims are listed. Procedures are set forth for selecting measurement instruments (including discussion of several specific instruments), and for collecting and analyzing data. Evaluation of classroom behavior and of instructional materials is outlined under the headings of student population, program scope, administrative considerations, methodology/classroom activities, teacher/student behavior, testing, materials, and facilities. Suggestions are offered for writing evaluation reports. Appendix A describes principal evaluation models, and Appendix B discusses two specific observation instruments. (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED176589

LANGUAGE IN EDUCATION: THEORY AND PRACTICE

19

Evaluating a Second Language Program

Gilbert A. Jarvis
Shirley J. Adams

FL 010823

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Center for Applied Linguistics
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Published by
Center for Applied Linguistics

Prepared by
ERIC Clearinghouse on Languages and Linguistics

Language in Education: Theory and Practice
Series ISBN: 87281-092-5

ISBN: 87281-105-0

September 1979
Copyright © 1979
By the Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

Printed in the U.S.A.

LANGUAGE IN EDUCATION: THEORY AND PRACTICE

ERIC (Educational Resources Information Center) is a nationwide network of information centers, each responsible for a given educational level or field of study. ERIC is supported by the National Institute of Education of the U.S. Department of Health, Education and Welfare. The basic objective of ERIC is to make current developments in educational research, instruction, and personnel preparation more readily accessible to educators and members of related professions.

ERIC/CLL. The ERIC Clearinghouse on Languages and Linguistics (ERIC/CLL), one of the specialized clearinghouses in the ERIC system, is operated by the Center for Applied Linguistics. ERIC/CLL is specifically responsible for the collection and dissemination of information in the general area of research and application in languages, linguistics, and language teaching and learning.

LANGUAGE IN EDUCATION: THEORY AND PRACTICE. In addition to processing information, ERIC/CLL is also involved in information synthesis and analysis. The Clearinghouse commissions recognized authorities in languages and linguistics to write analyses of the current issues in their areas of specialty. The resultant documents, intended for use by educators and researchers, are published under the title Language in Education: Theory and Practice.* The series includes practical guides for classroom teachers, extensive state-of-the-art papers, and selected bibliographies.

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to the American Council on the Teaching of Foreign Languages for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either ACTFL or NIE. This publication is not printed at the expense of the Federal Government.

This publication may be purchased directly from the Center for Applied Linguistics. It also will be announced in the ERIC monthly abstract journal Resources in Education (RIE) and will be available from the ERIC Document Reproduction Service, Computer Microfilm International Corp., P.O. Box 190, Arlington, VA 22210. See RIE for ordering information and ED number.

For further information on the ERIC system, ERIC/CLL, and Center/Clearinghouse publications, write to ERIC Clearinghouse on Languages and Linguistics, Center for Applied Linguistics, 1611 North Kent Street, Arlington, VA 22209.

From 1974 through 1977, all Clearinghouse publications appeared as the CAL-ERIC/CLL Series on Languages and Linguistics. Although more papers are being added to the original series, the majority of the ERIC/CLL information analysis products will be included in the Language in Education series.

Introduction

Many second language educators are uncomfortable with program evaluation. Yet, instructors at all levels of education and in all subject matter areas are now participating increasingly in the evaluation of their programs. In many instances they are being asked to plan and conduct formal reviews of their programs. Several factors seem to be converging to cause program evaluation to become an even more visible part of the educational process in the years ahead. Most notably, education is becoming more expensive each year. Its quality is becoming a matter of great public concern, and, at the same time, the body of knowledge dealing with educational evaluation is becoming increasingly sophisticated.

Language teachers in the United States are also experiencing the effects of several additional factors. The status of language study in the total context of our educational enterprise has not been favorable recently. At no level of education are language programs viewed as essential or basic to the quality of a person's education. We, as language teachers, have never generated a broad base of support. Public opinion polls repeatedly document our last-place status in public acceptance. Significant doubt about the quality of our programs is one of several factors that have contributed to this situation. ("They don't teach you the real language that everyone speaks.") Our history is not one of careful, dispassionate inquiry designed to determine optimum teaching strategies. We have, instead, a conspicuous history of conflicting methodologies and ideologies. We have, moreover, often promised greater proficiency than we could teach, and we have in many programs restricted our instruction to an elite segment of the school population. Factors such as these are responsible for a special type of vulnerability--

vulnerability that sometimes becomes manifest during evaluation procedures.

The notion of evaluation often creates feelings of doubt, apprehension, and antipathy among many language teachers. Evaluation has been threatening--both to their egos and to their careers. Unfortunately, evaluation in the mid-1970s has led more often to hostile decisions by administrators than to better education for everyone. In the worst cases, evaluation is reduced to a pretext for eliminating programs when the decision to eliminate has already been made though not promulgated. The classic case is that of the administrator who wants to reduce or eliminate the program and who structures the evaluation so that it accentuates weaknesses in the program.

The climate for evaluation has been further worsened by confusion between program evaluation and instructor evaluation. Program evaluation is not a euphemism for evaluating instructors. Instructors are only one component of the instructional process, and evaluation of instructional programs must involve all aspects of that process.

One particular type of evaluation, accreditation evaluation, is most familiar to teachers. Their experience with this type of evaluation does not, however, generally contribute to its perception as a means for improving instruction. Usually, it is viewed as an inevitable nuisance required by a remote and nebulous agency; it is an exercise to be done so that it can be forgotten for another ten years.

In an effort to accommodate the backgrounds of language educators who confront evaluation for the first time, the purposes of this paper are (1) to summarize important thinking and issues in evaluation as they relate to second language programs, (2) to describe procedures for evaluating a second language program, and (3) to describe procedures for analyzing evaluation data. The reader should recognize that because evaluation is a large and complex process, a brief paper such as this must omit many of its dimensions and aspects; moreover, every program is unique, and examples are of only limited utility. This document should therefore be considered a primer of second language program evaluation.

Evaluation activity must be premised on the conviction that conscientious, honest evaluation can lead to better programs and therefore to a more significant role for language educators in the total educational process. Evaluation should not be a distasteful or threatening task; it is a means to professional self-understanding and self-improvement. In the final analysis, however, it is only the behavior of educators that can transform these words from hollow echoes to more and better programs.

What Is Evaluation?

Evaluation can be an ambiguous term. On the one hand, we recognize it as an everyday activity in all our lives. We evaluate when we shop, when we turn on a television set and select a station, when we choose driving routes to work, or when we dress for a particular occasion. We may even be aware that each of us does not perform all these evaluations equally well; some are successful, but at other times we have to acknowledge that we have used bad judgment or made poor decisions. On the other hand, when we read the professional evaluation literature, evaluation is transformed into an apparently obscure process shrouded in technical jargon and concepts. In truth, we probably oversimplify our "lay" evaluation activity. Tyler et al. (1967) note that educational evaluation must always take into account the full complexity of the phenomenon:

> Educational programs are characterized by their purposes, their content, their environments, their methods, and the changes they bring about. Usually there are messages to be conveyed, relationships to be demonstrated, concepts to be symbolized, understandings and skills to be acquired. Evaluation is complex because each of the many characteristics requires separate attention. (1967:4-5)

The complexity of the teaching-learning process is intensified in the second language classroom. In our classrooms many dimensions of the process are in fact different or appear different from that of other subject matter areas. This is one of the most compelling reasons for

involvement--and therefore competence--in the evaluation process on the part of language educators.

An evaluator--whether a professional from outside or a teacher evaluating his or her own program--may be involved in such disparate tasks as judging the worth of a particular educational goal, determining whether a test does indeed measure the ability specified by a particular instructional objective, determining the prerequisite skills or sophistication for a particular instructional unit, or comparing a conventional one-hour-per-day time format to a one-month, all-day intensive program. Evaluation cannot be reduced to a simple procedure in which one follows a finite series of rigidly prescribed steps, nor is it uncomplicated conceptually. Definitions vary from author to author: "an assessment of merit," "determination of attainment of objectives," "procuring information to use in decision making," or "comparisons between alternate programs." All the definitions imply certain commonalities. They all require systematic efforts to define criteria and to obtain accurate information about the program characteristics. Evaluation must be characterized by what Cronbach and Suppes refer to as "disciplined inquiry":

> Disciplined inquiry has a quality that distinguishes it from other sources of opinion and belief. The disciplined inquiry is conducted and reported in such a way that the argument can be painstakingly examined. The report does not depend for its appeal on the eloquence of the writer or on any surface plausibility. The argument is not justified by anecdotes or casually assembled fragments of evidence . . . . The report of a disciplined inquiry has a texture that displays the raw materials entering the argument and the logical processes by which they were compressed and rearranged to make the conclusion credible . . . .
>
> Disciplined inquiry does not necessarily follow well-established formal procedures. Some of the most excellent inquiry is free ranging and speculative in its initial stages, trying what might seem to be bizarre combinations of ideas

and procedures, or restlessly casting about for ideas . . . . (1969:15-16, 18)

Cronbach and Suppes verbalize a very important principle for the teacher participating in an evaluation. It is crucial that the collection of information about the program be as free from bias as is possible. The information must be gathered dispassionately and organized systematically. If the principle of disciplined inquiry does not guide the process, evaluation in the true sense of the term cannot occur; instead, one creates propaganda. Thus, if students are interviewed during an evaluation, it is essential that they be representative of all the students in the program. If enrollment data are described, they must tell the "full story." Enrollment figures from the first day of class are not valid, for example, if 40 percent of the students drop the course before the end of one semester. Also, student achievement scores must include all scores or a representative sample.

Evaluation shares this quality of disciplined inquiry with research and program development. Earlier conceptualizations of evaluation were very narrow. It was made synonymous with measurement (testing), with professional judgment, or with comparisons between performance data and the objectives of the program. Recent definitions are more ecumenical and pluralistic, stressing both systematic procedures and judgment of worth. The conceptualization that guides this paper is similar to that of Scriven (1967) and virtually identical to that of Worthen and Sanders (1973): "Evaluation is the determination of the worth of a thing. It includes obtaining information for use in judging the worth of a program, product, procedure, or objective, or the potential utility of alternative approaches designed to attain specified objectives" (p. 19).

In a holistic approach, data must be collected on all variables or factors that may reflect the quality of a program--whether they be process variables or product variables. These data must be collected as systematically as possible. Once the data are collected, judgments of worth can then be made by asking whether the composite data indicate high or low quality. Similarly, judgments can be made about particular components or dimensions of the program. Should the particular component be retained, modified, or replaced by an alternative arrangement?

This approach assumes, furthermore, that a program cannot exist in a vacuum. A program is influenced by and itself influences various persons or groups with differing needs and points of view. Procedures for carrying out an evaluation must take into account the nature of the particular program. An evaluator must be analytical and circumspect.

[For the reader interested in the full range of conceptualization of evaluation, a brief summary of general models is given in Appendix A.]

Purposes of Evaluation

One useful way to conceptualize the purposes of evaluation is to dichotomize them into *formative* and *summative* evaluation. The purpose of formative evaluation is to improve the instruction. It asks, in effect, about the current status of the program so that it can be made better. It is evaluation that is carried out during the development, implementation, and operation of the program. Summative evaluation is terminal evaluation of a program that is already operational. Its purpose is to make judgments about the program's worth. One can also use it to determine which evaluative labels to place upon the program ("Great," "Good," "Mediocre," etc.). Ultimately, summative evaluation is tied to decisions about support and continuation of a program.

These two purposes are not mutually exclusive. A summative evaluation can subsequently be used for formative purposes. Similarly, a formative evaluation at a given moment can lead to answering such questions as whether or not enough progress has been made in the development of a program. It should be noted, however, that some procedures are not equally suited for the two types of evaluation. In a formative evaluation it is important to obtain day-to-day feedback on specific aspects of the instruction. This information is, however, often fragmentary and of limited use in a summative evaluation. Conversely, an assessment of the impact of a new program on the image of language study in a particular school may be very important in a summative evaluation but is inappropriate for a formative purpose.

Formative and summative evaluators often do not "behave" in the same way. A formative evaluator's procedures may be more partisan than a summative evaluator's approach. The summative evaluator must be objective and circumspect. A formative evaluator can use shortcuts, small samples, and intuition in an effort to improve the program.

There is a need for more formative evaluation in language programs. Rarely do we initiate evaluation for the purpose of improving our programs. Even in light of the very heavy loads of most instructors, evaluation is still far too important an activity to be so neglected.

Defining the Program

At first glance, defining a program may seem to be an unnecessary task--one that is so obvious that no effort is necessary. In many schools and colleges, however, it may be very difficult to make decisions about the scope of an evaluation.

Consider the following example of a fictitious high school (Central High). The staff at Central High has decided to undertake an evaluation of its second language program. The school houses grades eight through twelve, though eighth and ninth grade students are classified as junior high students. Four major languages are taught--French, Spanish, German, and Latin. The program offers four years of study in each language. In addition, the assistant principal teaches eight students Italian twice a week, though no credit is given. Most students begin studying a language in the ninth grade, but Central High recently implemented an exploratory language program for junior high students, as well as an immersion program for qualified students, which is offered during the summer by temporary staff hired expressly for that program. Furthermore, the social studies department offers courses on French and Spanish history, and the English department offers courses in European literature. A language lab is shared with the business department, which uses it to teach shorthand.

Before beginning evaluation procedures, the staff at Central High has some decisions to make. Will the evalu-

ation cover all the courses in all the languages offered, or is Latin to be excluded from this evaluation? Should the exploratory program offered for junior high students be included, or is it part of the junior high language arts program? Should the summer immersion program be included in the evaluation, or is it considered a special program? Because the special courses offered in social studies and English are often taken in conjunction with the language studied, should these courses also be included in the evaluation?

Questions such as these, though sometimes very difficult, must be answered to the satisfaction of all participants prior to the start of an evaluation. Each situation is unique, and there are no firm guidelines to follow. The kinds of decisions to be made on the basis of the evaluation, the resources and time available, and the political implications of the various alternatives must all be taken into account.

Once the scope of the program to be evaluated is determined, an accurate description of who and what are involved in it can be prepared. This descriptive document serves two purposes: it brings together all who are interested in the evaluation, and it clarifies in the minds of the evaluators exactly what the program's characteristics are. This document can be brief, but it should include the following:

1. Demographic information about the staff (e.g., backgrounds, degrees, experience, etc.)

2. Demographic information about the student population

3. Description of the course offerings and enrollments

4. Relevant historical information (e.g., Is it a long-standing program or relatively new?)

5. Description of facilities

6. Pertinent information about the role of the program in the total offerings of the school

Who Is to Do the Evaluation?

The backgrounds of those who conduct an evaluation determine to some degree the quality and nature of the process. An outside evaluation expert will conduct a methodologically sound evaluation but may miss important information specific to our field. Such evaluations are also often restricted in scope and time because of the expense involved. Evaluation by a foreign language education expert may be methodologically less elegant; however, the specific competence in our field is a compensating factor. This type of evaluation usually also operates under scope and time restrictions for reasons of cost. Evaluation solely by the staff of a program introduces various kinds of biases and school politics. It may be better than no evaluation at all, but it is the least attractive option.

The most feasible solution to these limitations seems to be a combination of talent. A carefully done self-study by the program's staff followed by an evaluation by an outside expert in language education provides many advantages. (Ideally, the expert is also experienced in evaluation.) It capitalizes on the local, relevant skills, yet provides the conscience, fresh outlook, and expertise of an outsider. Such a procedure facilitates the work of the outside expert, frequently permitting that person to accomplish his or her evaluation with no more than two days spent visiting the school. Thus, expense is reduced and quality maintained.

Evaluating the Goals and Objectives of the Program

The educational aims of a program are a crucial initial consideration in evaluating any type of instruction. To accept goals without questioning their validity or desirability is to commit a serious error that reduces the remainder of the evaluation to an exercise in irrelevance. The goals and objectives are important from three perspectives:

1. It is largely in terms of goals that one must judge the classroom behavior of learners and teachers.

2. There are more and less desirable ways for formulating goals. The process by which they are formulated and utilized must itself be judged.

3. The content of the goals and objectives must be judged. Is it consistent with the needs and nature of society? Is the content consistent with the most up-to-date knowledge of the teaching/learning process? Is the content consistent with the nature of the students?

Goals cannot be considered an optional component of a second language program. They are essential. Their inclusion is not the result of any educational fad, nor is it a simple matter of responding to societal trends like accountability or specificity. Some would claim, for example, that we are in an age of specificity and accountability and that the use of goals and objectives in educational programs is merely one manifestation of that phenomenon. Some instructors, moreover, have experienced an administrator's fiat that their curriculum must be designed around objectives of a particular type (usually behavioral) "beginning next Monday." Their attitudes are predictably hostile. This is unfortunate, for education is inherently purposeful. Students have purposes or aims, and so do teachers. These aims are inevitably present whenever two persons come together in teacher and student roles. To the degree that their respective purposes coincide, the educational process will be enhanced. But even when the purposes or objectives do not coincide or are not verbalized, they are nevertheless present.

Statements of objectives serve purposes beyond clarifying the intent of their formulator: they function as a communication device among all groups involved in the educational process, including teachers, administrators, parents, and other interested parties. There are many ways to formulate statements of aims. One important distinction, however, relates to their level of generality. Statements of aims can be in very broad, general terms or, at the other extreme, in very specific, concrete behavioral terms. The former are often referred to as statements of "philosophy" in secondary schools and "mission" statements in colleges and universities. They represent the most succinct statements of the raison

d'être for the program; moreover, they represent the point of departure or basis for the formulation of all the specific objectives. From a statement of goals (the term "goals" is frequently applied to this most general statement of aims), one can proceed through increasingly specific statements of aims in regard to abilities or knowledge, ultimately reaching very explicit statements of behavioral objectives (the term "objective" being conventionally applied to specific statements of aims). An objective may, for example, lead to a specific classroom activity on a particular day. The number of stages from the goal statement to the most specific objectives is not nearly so important as their continuity. It is imperative that the specific objectives that guide day-to-day behavior add up to the goal of the program.

As an example of this progressive differentiation of aims, one might envision a program consisting of goals, sub-goals, end-of-course objectives, and learning objectives. The broad goal statement is broken down into several sub-goals, which, in turn, are broken down into objectives for the end of a semester, quarter, grading period, or even an academic year. These end-of-term objectives are further categorized into a large number of learning objectives, which guide decisions about the moment-to-moment behavior in the classroom.

Schematically, the process is represented by Figure 1. Sample content for the lettered boxes is given below. As mentioned earlier, such statements must be developed locally for each program and cannot be "copied" from any source.)

A. (Sample Goal Statement)

Each and every moment of one's life can be viewed as a continuing effort to communicate or to interact with one's total environment. Students should therefore have the opportunity to enhance their knowledge and ability to communicate and interact with the world that surrounds them. This ability implies understanding the communicative process, recognition and acceptance of new patterns of thought, awareness of one's self and how individuals differ from one another, a positive self-concept, an adaptability and

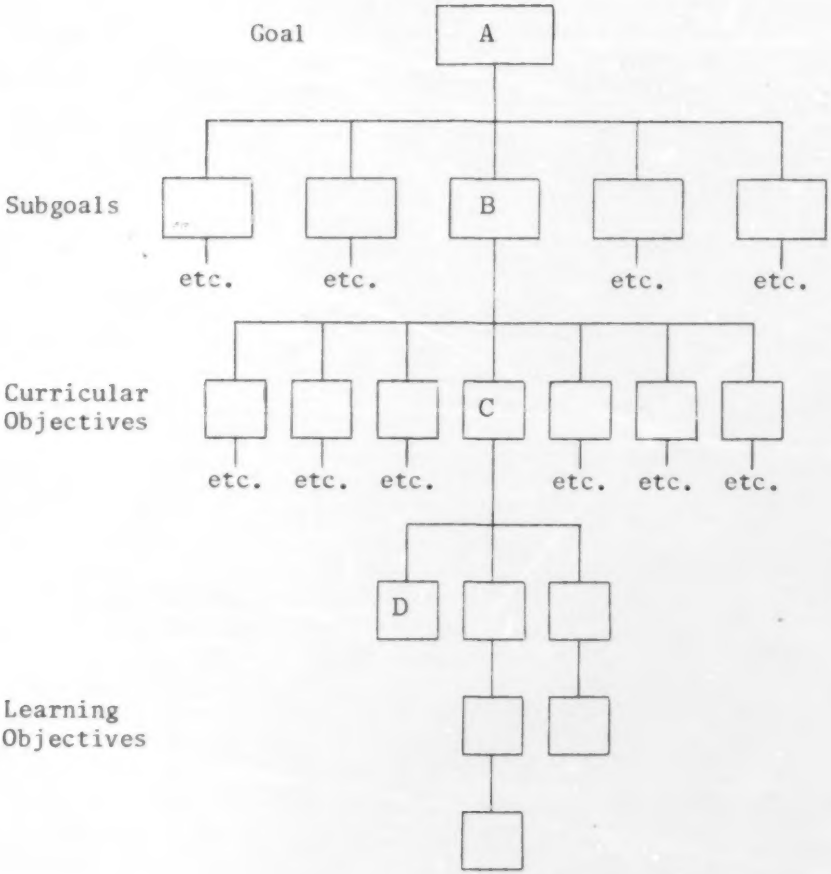

Figure 1

flexibility in the face of the unfamiliar and change, and diverse intellectual skills. The study of a foreign language can lead to these outcomes.

Note that the goal statement is very broad or general ("philosophical," to some). It is also concise in that it says a great deal about the aims of the program in relatively few words. The content delineates the contribution that the study of a language can make to the total education of a person in the late twentieth century. Goal statements relate our subject matter to the fundamental purposes of all education.

B. (Sample Sub-Goal Statement)

Students will develop the ability to communicate orally in the language.

Sub-goals are one step more specific. Each of them represents a segment of the learning implied in the goal statement. It is more specific than the goal: it focuses on oral skills. It is not sufficiently specific, however, for making hour-to-hour decisions about classroom behavior.

C. (Sample Curricular Objective)

Students will be able to ask and answer conversation-type questions about the weather in a way that is understood and does not irritate native speakers.

Curricular objectives are one step less general and more concrete. These end-of-term or end-of-course objectives are usually phrased in behavioral terms, though the behavior specified is an amalgamation of specific behaviors.

D. (Sample Learning Objective)

Students will be able to use the irregular verb faire with 80 percent accuracy in idiomatic expressions to indicate it is warm, cool, cold, windy, and nice in oral responses to questions.

Learning objectives are very specific. They describe behavior that students will exhibit within one class period or perhaps a few days. They indicate student behavior in regard to particular structures and vocabulary. It is important, however, that each specific learning objective be a particle of the overall goal. Too often, they are created with no consideration of the overall purpose of the program.

Thus, the goal refinement process must be viewed (and must be judged) as a process of progressive differentiation. Goal statements provide direction for a program and engender specific objectives that guide moment-to-moment classroom behavior. Any given classroom technique

or strategy can only be judged by reference to an objective. One cannot ask in the abstract, for example, whether a translation is good or bad or whether a particular drill should be longer or shorter. Such questions can be answered only within the context of the initial learning objective.

The content of the goals and objectives is never easily decided. Such decisions are always value judgments made about inherently complex matters. The needs and nature of society must be considered and interpreted. Moreover, it is a society of tomorrow that must be envisioned, for our students are going to spend the major part of their lives in the twenty-first century.

The nature of the learner, including inherent limits on what can be learned in a given amount of time, must be considered. Language educators have in the recent past been guilty of violating this fundamental requirement. In many programs we have specified unattainable goals. Some of us have unrealistically said, for example, that students would achieve what amounts to a form of bilingualism by the end of four semesters of language study. We have also tended to expect unrealistic ability in the writing skill--particularly in consideration of its difficulty and importance.

The body of knowledge about the second language teaching/learning process must be considered in formulating or judging goal statements. We have, for example, become increasingly aware of the importance of genuine communicative or meaningful use of the language rather than purely manipulative use. In the past, too many programs had implicit--if not explicit--goals that stopped short of communicative use. As a result, a large number of students were very skilled in doing all forms of drill and exercise but could not ask directions to a hotel or read a menu. The transfer from drills and exercise did not occur automatically. One of the important contrasts between today's goals and those of a decade ago relates to this recognition. Many evaluators today find goal statements that are slightly anachronistic because of this excessive emphasis upon rote learning and drill.

A growing number of goal statements reflect current interest in non-language skill outcomes. Cross-cultural

understanding, insights into the communication process, adaptability in the face of the unfamiliar, and mental dexterity are increasingly specified as aims of the study of a second language.

Guideline Questions for Judging the Statement of Aims

Are the statements of aims clear and well defined?

Do the specific learning objectives have a direct and clear link to the broad goal statements?

Are the goals realistic? Are the objectives appropriate for the level of instruction?

Are the goals consistent with the needs and interests of the students? Of the staff? Of the governing bodies of the school? Of the community or area? Of the contemporary world?

Are the goals consistent with learning theory? Are they organized and sequenced for efficiency?

Are the goals compatible (e.g., not emphasizing both accuracy of pronunciation and ability to communicate from the first days of instruction)?

Selecting Instruments, Collecting Data, and Analyzing Results

In many phases of an evaluation it is necessary to collect information from relatively large numbers of persons (e.g., students in the program, other students, community members, faculty members). It is usually most efficient to do this with measurement instruments (tests, questionnaires, etc.). It follows, therefore, that the quality of these instruments must be carefully considered. The wise selection (or creation) of instruments, the careful collection of data, and correct choice of procedures for analyzing the data are essential to a high quality evaluation. A plan for all data collection and analysis should be made prior to starting an evaluation. This plan will assure that all information relevant to the quality of a program will be

collected. It should be reemphasized that only after complete and accurate descriptive information has been collected can judgments be made about the worth of the program.

Using Existing Information

Before selecting or constructing instruments to obtain the desired data, the evaluator should determine whether or not the information might already be available from other sources such as instructors' records, school records and files, counselors' or advisors' files, records of other personnel (e.g., admissions office, extra-curricular activity advisor, coaches), or records held by parents or students themselves. In many situations, the above sources will be of limited value because the information may not directly address the questions being asked in the evaluation. An evaluator may, for example, want to determine what percentage of students have studied a second foreign language, and the high school records may not indicate languages studied in junior high school. Another concern is the accuracy of the data. Because the evaluator has had no control over the way in which the information was collected, its reliability may be questionable and the data therefore unusable.

Selecting Instruments

When the information needed cannot be obtained from available records, it is necessary to select or construct instruments to obtain the data. The most important consideration in selecting instruments is whether they measure adequately what they are intended to measure. If a test is intended to measure proficiency in the target language, does it indeed measure the appropriate language skills, does it use the important vocabulary and structures, and is it at an appropriate level of difficulty? If the instrument is a questionnaire to determine student backgrounds or attitudes toward language study and toward the instruction, is it written with clarity and at an appropriate level of sophistication? (For instance, high school students will not understand terms like "ethnocentrism.") Other considerations are

1. Reliability: If the test could be re-administered (after all the effects of having responded once before had been magically erased from the minds of those responding), would it yield similar results? Common threats to reliability from within the instrument are ambiguous questions and poor quality reproduction.

2. Ease of use: Is the instrument relatively easy to administer and summarize or score?

3. Time and resources: Can the instrument be used within the time blocks available (e.g., a class period)? Does it have to be purchased? If so, is the cost reasonable for the benefit to be gained?

4. Population: Is the instrument appropriate for the group with which it will be used?

5. Form of the resulting data: Does the instrument yield objective data? If not, is there provision for maintaining the quality of subjectively collected data? If several persons will collect the data, is there control for the differences?

Types of Instruments

It is impossible in a paper such as this to provide great detail about types of instruments. This is one area where many language teachers may wish to consult measurement experts or psychometric textbooks, if a member of the staff is not already familiar with use of the particular instrument type.

Once a decision is made as to what is to be assessed (achievement, performance, attitudes, interaction among persons), the type of instrument can be determined. One useful way for categorizing instruments is in terms of four broad types (TenBrink, 1974): observation, inquiry, analysis, and testing.

Observation may include such items as anecdotal records, checklists, rating scales, and rankings. These methods

may or may not be very objective. The quality depends on how well the criteria are defined for each instrument.

The category of inquiry includes questionnaires, interviews, and--for certain types of evaluation--sociometric instruments and projective techniques. Questionnaires and interviews are particularly useful for identifying attitudes held by language students--attitudes that are pertinent to the quality of the program.

Some language educators have successfully used the Foreign Language Attitude Questionnaire (FLAQ) (Jakobovits, 1970). It is intended for all ages and has two forms--one designed for students who are studying or have studied a foreign language and the other for those who have not studied a language. The questionnaire asks about the students' language background, their attitudes toward foreign language study, and their reasons for study. The format is multiple-choice with an opportunity for additional comments. It may be most useful as an example or model, for most language educators would be most satisfied with construction of an instrument designed for their own particular curriculum.

Analysis, as defined by TenBrink, includes the techniques of content analysis and interaction analysis. Content analysis is a counting procedure, whereby written or spoken communications are analyzed for the presence or absence of certain characteristics (1974:146). Interaction analysis is useful for obtaining information on group participation, individual student interaction with the group, and instructor interaction with the class. It can provide descriptive information about the percentage of time spent by the instructor and students using the target language versus using the native language, or teacher talk versus student talk, or any other categories of behavior that are deemed relevant. [A more extensive discussion of these systems appears in Appendix B.]

The fourth category, testing, is the one most often used for evaluation purposes. Achievement tests and diagnostic/prognostic tests are available commercially. Diagnostic tests include the Modern Language Aptitude Test (MLAT) (Carroll and Sapon, 1959), the Elementary MLAT (Carroll and Sapon, 1967) and the Pimsleur Language Apti-

tude Battery (LAB) (Pimsleur, 1966). The MLAT has been widely used for many years. It is designed for English-speaking persons from grade 9 through adult. It has two forms--a long form requiring 60-75 minutes to administer, and a short form requiring approximately 40 minutes. The short form simply omits parts 1 and 2 of the long form (and has nearly as good validity). The total test has 5 parts:

1. Number learning: Students learn and then are tested on numbers in a new language.

2. Phonetic script: Students learn sound-symbol correspondences and are then asked to select the correct transcription for spoken words.

3. Spelling clues: Students select synonyms for coded English words.

4. Words in sentences: Students must manipulate various grammatical concepts without the use of any grammatical terminology.

5. Paired associates: Students memorize vocabulary in a new language.

The Elementary Modern Language Aptitude Test is similar to the MLAT but has tasks that have been simplified to make it appropriate for younger students down to grade 3. It requires about 60 minutes to administer. (There is no short form.)

The Pimsleur LAB takes approximately 60 minutes. It is intended for English-speaking students in grades 6-12. It consists of six parts, which the test defines as the components of aptitude: Grade-Point Average, Interest, Vocabulary, Language Analysis, Sound Discrimination, and Sound-Symbol Association.

The principal role for aptitude tests in a language classroom is as a diagnostic device. In an evaluation, an aptitude test can provide information about the student population and therefore about the match between student characteristics and the type of curriculum.

Achievement tests include the Common Concepts Foreign Language Test (California Test Bureau, 1962), the MLA Cooperative Foreign Language Tests (Educational Testing Service, 1963), and the Pimsleur Modern Foreign Language Proficiency Tests (Pimsleur, 1967).

Appropriate for all grades, the Common Concepts Foreign Language Test is about 40 minutes in length and tests listening comprehension in French, German, and Spanish. Students hear sentences in the foreign language and indicate their understanding of what they have heard by selecting from sets of four colored pictures the ones that have been described.

The MLA Cooperative Foreign Language Tests (French, German, Italian, Russian, and Spanish) measure listening, speaking, reading, and writing skills. One form is appropriate for levels one and two, the other for levels three and four.

The Pimsleur Modern Foreign Language Proficiency Tests are available in French, German, and Spanish for levels one and two. They also test the four skills.

Care must be exercised when choosing standardized tests. Klein (1971) points out that standardized tests often have such limitations as questionable validity, poor overlap between program and test objectives, inappropriate instructions and directions, and confusing designs and formats. The overlap between vocabulary used on these tests and vocabulary taught in most U.S. language programs is discouragingly small. In most situations, a language staff can create an instrument that provides data that are more trustworthy than the results from any commercially available test. Test construction, however, requires considerable time, effort, and expertise. If a test is to be constructed, the procedures should reflect standard procedures used by professional test makers. It should, first of all, incorporate the best judgment of all faculty members. This test should then be given experimentally to a group of students and should be revised on the basis of the way the test functions. Only then is it ready for use in the evaluation.

It must be remembered that data from all instruments are simply a description of outcomes and characteristics of

the program and its participants. The choice of instruments is therefore synonymous with the question, What kinds of information will be necessary in order to make judgments about the quality of the program?

Collecting the Data

After the appropriate instruments have been created or selected for obtaining the desired information, the procedures for collecting the data must be determined. Collecting data is an important step requiring careful planning. It can involve delicate interpersonal relations, because its success often depends on the cooperation and assistance of persons who are not directly connected with the evaluation. If a "data collector" inadvertently makes a mistake, it could affect the validity of the data. Thus, careful training and specific instructions are essential considerations for an evaluator. Law and Bronson (1977) suggest the following steps:

1. Make the necessary arrangements with the school(s) and the personnel who are to be involved.

2. Decide who will collect the data.

3. Make the arrangements for any training that is needed.

4. Schedule the collection within the time allotted for this phase of the evaluation.

5. Monitor the total data collection process.

An important consideration is whether the data collection is to be planned and carried out by someone within the program or by someone from outside. If it is done by personnel within the program, they should be aware of two potential sources of bias. First, they may have a vested interest in the program and may focus excessive attention on its successful aspects while de-emphasizing troublesome elements. Second, they may be so conscious of and attentive to the program objectives that they overlook secondary effects that are equally important.

Analyzing the Data

It would be very convenient if data that have been collected could be fed into a computer, which would then feed back a printout with both a summary and an interpretation of the information. Unfortunately, neither the summarization nor the subsequent judgment is that simple.

Summary of data requires many decisions. Counting frequencies and percentages is mechanical; however, categories may be modified (especially by being combined or eliminated). One may, for example, elect to combine "agree" and "strongly agree" categories in a scale because few respondents have chosen the strongly agree or strongly disagree categories. It may also be important to emphasize or highlight certain aspects of the data. When this is done for a valid purpose, it is of course justified. One must recognize, however, that there is a very fine line between emphasis and distortion.

It may be useful with some types of data to ask whether a pattern or difference is due to particular characteristics or features of a program or due to chance variation. Tests of statistical significance (chosen by someone with relevant competence) will provide that answer.

A distinction must be made between statistical significance and practical significance. To determine statistical significance, one calculates the probability that a given event could have occurred by chance alone. If the probability of occurrence by chance is small, the researcher concludes that the results are due to non-chance factors or to the condition or program under investigation.

Statistical significance is sensitive to the number of people, items, or events involved. The larger the sample size, the more likely one will be able to rule out chance as accounting for a particular difference or pattern. Statistical significance is only part of the total picture. Because important decisions must be made on the basis of an evaluation, it is often necessary to show practical significance as well. Program evaluators might have determined, for example, that a group of students who used the language laboratory were superior in listening skill to other students who experienced the same

instruction but who did not use the laboratory. Let us assume that with the help of a statistician it has been determined that the few points of advantage for the laboratory students are statistically significant. It is entirely possible, however, that the staff (or administrators) might judge that the slightly better listening ability is not worth the high cost in dollars of maintaining the lab.

Thus, value judgments must be made "on the bottom line" regardless of the sophistication of the data. The numbers function to show variation, but they cannot tell whether or not the variation is desirable. The evaluators must look at the numbers and then judge whether they are "good" or "bad." Such value judgments can never be made in the abstract.

Evaluating Classroom Behavior

The mass of classroom behavior by instructors and students is extremely large, diverse, and complex. There is extensive evidence that we do not understand well the phenomena involved. Research on how people learn in general and how they learn foreign languages in particular is in its infancy. Teaching effectiveness is the most researched area in all of education, and yet there are pitifully few generalizations one can make about factors that contribute universally to success. In order to evaluate, one must either resort to a single global impression of the quality of the instruction or break the many dimensions of the behavior into manageable units. Yet, at the same time, each small unit or segment of activity must not be separated from the context in which it occurs. Out of context it becomes uninterpretable. To ask, for example, whether or not a particular homework assignment was appropriate requires consideration of many factors or variables. What is the overall objective? What is the specific purpose of the activity? What preceded it? What will follow it? What is the level of instruction? What are the characteristics of the learner(s)? What is the instructor's teaching style? Such a list of questions could go on for pages, but nevertheless must be asked intuitively whenever one is judging any segment of classroom behavior or materials.

As demanding and difficult as the evaluation of instruction is, it must be judged. It is in the classroom that the impact of all variables becomes real. As with other phases of evaluation, the first stage requires an accurate perception and description of the pertinent phenomena. Only then can the second stage--judgment of worth--take place. As discussed elsewhere in this paper, behavioral observation systems (and other instruments) are one means of increasing the objectivity of the description. (There remains, of course, subjectivity in the choice of categories of behavior that are to be observed.) It is not possible, however, to observe systematically all dimensions of instruction. Frequently, the most defensible evaluation is a rating of a dimension of the instructional activity and materials. On the basis of many criteria which fall under the rubric of "experience and knowledge," the evaluator observes carefully and dispassionately and then makes a judgment.

Listed below is a series of questions designed to guide observation and judgment of instructional behavior and materials. There are no single "right" or "best" answers to them, and the list is not all-inclusive. For some of the questions there would be widespread agreement on a most desirable response; for others, there would be little agreement; for still others, the most desirable response depends on the particular teaching/learning situation. Each draws attention to a particular aspect of instruction so that the evaluator can make a judgment about it. Each is also conducive to a "why" or "why not" follow-up question.

A. The Student Population

1. Is the program open to all students, or is it restricted to certain types or categories? Are the restrictions valid?

2. Does the program attract a sufficient proportion of those students who are eligible to enroll?

3. Are students well informed about the nature of the program, its goals, and the benefits of language study?

B. Program Scope

1. Is the scheduling of language classes consistent with the program goals and the interests of the faculty?

2. Can students move comfortably from one course or one level to the next?

3. Is the length (number of courses or years) of the program appropriate? Is the breadth of offerings adequate?

4. Is the overall sequence of content logical? Is it consistent with the program objectives and today's knowledge?

C. Administrative Considerations

1. Is there good communication with the rest of the school or university? With those responsible for the advising and counseling of students? With the administration? With other faculty members?

2. Is the ratio of students to faculty acceptable?

3. Is there an adequate amount of extracurricular activity relating to language study?

4. Are community resources utilized in the program (e.g., native speakers, companies with international contacts, museums, etc.)?

D. Methodology/Classroom Activities

1. Is the methodology consistent with the goals of the program? With the specific objectives of any particular moment?

2. Is the methodology logically consistent?

3. Is the moment-to-moment sequence of activities logical (e.g., simple to complex, beginning with the known, leading to genuine communicative ability)?

4. Are the steps from one activity or segment to the next of the right size (i.e., do they challenge the students while not being too difficult)?

5. Are sufficient examples and models given?

6. Are presentations clear and interesting?

7. Is the classroom pace appropriate?

8. Are correction strategies and behavior wisely used?

9. Is there adequate opportunity for practice of the language?

10. Do students work individually, in small groups, or in large groups according to the nature of the task and its purpose?

E. Teacher/Student Behavior

1. Are the instructors fully aware of the goals?

2. Is the teacher role consistent with the philosophy of the school or department? Are instructors comfortable in their role?

3. Do instructors have an opportunity to interact with colleagues? To visit the classes of other instructors?

4. Is the moment-to-moment teaching behavior valid?

5. Do the instructors consider student characteristics such as age, aptitude, motivation, and interests in all their interaction with them?

6. If student attitudes have not been surveyed by formal instruments, does their behavior reveal their overall attitudes toward the program? Toward specific components of the program? Have these attitudes been taken into account in improving the program?

7. Is feedback given to students about their performance? Does it take into account individual differences?

F. Testing

1. Is testing valid in terms of program goals? Does it measure genuine communicative or meaningful use of the language?

2. Does testing use efficient established procedures and item types?

3. Does the testing have face validity (e.g., do students perceive it as fair)?

G. Materials

1. Are the instructional materials consistent with the goals? Is the basic text supplemented by readers, workbooks, or other materials?

2. Are audiovisual aids available and used effectively?

3. Are the materials authentic? Do they create intercultural awareness without reinforcing stereotypes? Do they represent all segments of society and all cultures where the target language is spoken?

4. Are teacher-made materials of good quality?

H. Facilities

1. Are the facilities adequate? Do they meet basic needs (e.g., space, light, ventilation, etc.)? Are they conducive to teaching and learning?

2. Is the library adequate? Does the staff have a professional library?

3. Are there adequate support services (e.g., clerical, aides, etc.)?

Honest and accurate answers to such questions form the basis for judgments about the value or worth of instructional components. They are probably the most important evaluative activity in judging the process (as opposed to the products) of education.

Writing the Evaluation Report

The preparation of a final document that communicates to all concerned the findings of the evaluation is an important last step in the evaluation process. Perceptions presented orally are very susceptible to varied interpretation. Furthermore, considerable clarification of the ideas occurs during the preparation of the document.

The format is a matter of preference and style. The content will vary depending upon the local conditions. In one way or another, it will include what was observed and any recommendations that may be made. In its simplest form, where there has been a self-study, it will merely endorse the self-study.

The following outline reflects one possible "table of contents" for an evaluation report.

I. Objectives of the Evaluation

- A. Rationale for evaluating
- B. Audience
- C. Decisions that may be anticipated

II. Description of the Program

- A. Educational philosophy
- B. Goals and objectives
- C. Staff
- D. Instructional procedures/methodology
- E. Content
- F. Student population
- G. Community setting
- H. Facilities

III. Program Outcomes

- A. Student achievement

B. Attitudes
C. Side effects
D. Costs

IV. Judgments about Program

A. Value of outcomes
B. Strengths
C. Weaknesses
D. Recommendations

Conclusion

Obviously, any evaluation report that is simply filed after being completed has demanded far more effort and time than can be justified. An evaluation that is utilized to improve instruction is not only beneficial to everyone associated with a program but is also very satisfying to those who participated.

There is a critical need to do more evaluation of our language programs. It is especially important that we conduct evaluations at "non-crisis" times--times when our programs are not under direct attack. Such evaluations should be solely for the purpose of improving the programs.

The most useful evaluations, moreover, are those that are designed and planned at the local level. Such evaluations benefit from general guidance, but because each situation is unique, precise specifications cannot be made. It has been the intent of this paper to provide the guidance needed for successful evaluations. One source of consolation for anyone facing the many decisions that have to be made in doing an evaluation is the fact that there are always multiple "good" decisions. Evaluation is not an activity in which one must follow a prescribed set of procedures. It comprises, in the final analysis, all activities in which one carefully gathers information and then makes judgments of worth or quality. This model is well suited to all professional activity.

Appendix A

Principal Evaluation Models

Because of the complexity of educational evaluation, it is not surprising that many evaluation experts have turned to the use of symbolic models to clarify evaluation. Like all symbolic models, they achieve manageability by reducing the mass of behavior involved in evaluating to a series of abstractions. The most frequently discussed (and presumably utilized) models, as identified by House (1978) and as described throughout the professional literature, are

1. *Systems Analysis*
   One begins with measures of results of the program (output measures) and attempts to relate them to characteristics of the program. One asks whether or not all dimensions and components of the system are functioning effectively. Variations in the results of the program (e.g., student learning or enrollments) are traced to changes in the program characteristics (e.g., materials used, methodology, class time).

2. *Behavioral Objective Attainment*
   The objectives of a program are delineated in very specific terms of student performance. To determine the extent to which these objectives are being attained, students are tested, and the results are compared with the objectives. The extent to which student performance matches that specified in the objectives is an index of the program quality.

3. *Decision-Making Model*
   One begins with the decisions that are to be made on the basis of the evaluation. These decisions guide the nature of the evaluation. The evaluation supplies information that is relevant to the decision. Thus, if one were to make decisions about the student population that was to study a language, variables such as age, grade or year, language aptitude, college-bound or not, major, and other grades might receive special attention.

4. *Goal-Free Evaluation*
   In an effort to reduce the effects of bias in evaluation, the intents of the program developers or decision

makers are not revealed to the evaluator. The evaluator must therefore search for all outcomes of the program. The results of such evaluation can be used by the developers to improve the program and by consumers (students) to accept or reject it. In some ways this model is similar to the Consumer Reports approach to evaluation.

5. *Art Criticism Model*
The educational critic operates in the same manner and tradition as the art or literary critic. A major assumption is that the evaluator has become skilled by his or her training and experience to judge the important aspects of a program. This model is frequently used in second language programs. Usually, an expert such as a foreign language education professor is brought in from the outside to spend several days on campus or at the school to make judgments about the program. Much of this person's expertise lies in how well he or she can intuitively implement principles of judgment.

6. *Accreditation*
This frequently used form of evaluation usually involves visits by teams of colleagues from other schools. Principles of judgment are usually spelled out in checklists of evaluative criteria. The local staff have previously collected and analyzed information in an extensive self-study. In many ways, the visitors function as a conscience for the self-study.

House also identifies as other major models the Adversary Model, in which the pros and cons of a program are argued in a manner not unlike a trial by jury, and the Transaction Model, in which the educational processes themselves are studied. Neither of these models seems to be in widespread use in foreign language education.

Appendix B

Observation Instruments

Observation instruments can be very useful in describing classroom behavior. Essentially, they consist of a system of categories of behavior into which instances of the particular behaviors are classified as they take place. The instrument may be used to record all instances of the categories or a sample of them. Sampling is usually done via a time sample in which occurrences of behaviors are recorded at a regular interval (e.g., every 10 seconds or every 15 seconds). The behaviors are then summarized in such forms as matrices, percentages, or simple frequency counts.

Observation instruments can be very useful in gathering accurate information about what occurs in classrooms. The technique probably, however, holds more potential than has been realized so far. Instructors should therefore feel free to modify existing system or to create their own to meet their own needs.

There are several instruments available for coding classroom behavior. Flanders (1960) developed one of the first and most popular instruments. The Flanders Interaction Analysis Categories (FIAC) contain seven categories for coding teacher verbal behavior and three categories for judging pupil verbal behavior:

TEACHER TALK

Indirect Influence

1. *Accepts feeling:* accepts and clarifies the feeling tone of the students in a non-threatening manner. Feelings may be positive or negative. Predicting and recalling feelings are included.
2. *Praises or encourages:* praises or encourages student action or behavior. Jokes that release tension, not at the expense of another individual, nodding head or saying "uh huh?" or "go on" are included.
3. *Accepts or uses ideas of student:* clarifying, building, or developing ideas or suggestions by a student. As teacher brings more of his own ideas into play, shift to category five.
4. *Asks questions:* asking a question about content or procedure with the intent that a student answers.

Direct Influence

5. *Lectures:* giving facts or opinions about content or procedure; expressing his own ideas; asking rhetorical questions.

6. *Gives directions:* directions, commands, or orders with which a student is expected to comply.

7. *Criticizes or justifies authority:* statements intended to change student behavior from nonacceptable to acceptable pattern; bawling someone out; stating why the teacher is doing what he is doing, extreme self-reference.

STUDENT TALK

8. *Student talk-response:* talk by students in response to teacher. Teacher initiates the contact or solicits student statement.

9. *Student talk-initiation:* talk by students, which they initiate. If "calling on" student is only to indicate who may talk next, observer must decide whether student wanted to talk. If he did, use this category.

10. *Silence or confusion:* pauses, short periods of silence, and periods of confusion in which communication cannot be understood by the observer.

The FIAC has been modified and extended by various researchers in an attempt to adapt the instrument to varying philosophies and subject matter areas. Moskowitz (1968) created an adaptation of the FIAC for the foreign language classroom. Her Foreign Language interaction system (FLint) includes, in addition to Flanders' categories, the following items:

1. the teacher
   - -jokes
   - -repeats student ideas verbatim
   - -corrects without criticism
   - -directs a pattern drill
   - -criticizes student behavior
   - -criticizes student responses
2. silence

3. confusion
4. laughter
5. English

Moskowitz also added a foreign language I/D ratio (indirect/direct), an English I/D ratio, and the F/E ratio (ratio of foreign language to English) for the total lesson.

Jarvis (1968) also developed a system for observing foreign language classroom behavior. His instrument classifies behaviors in terms of language skill acquisition consequences of the behaviors. The instrument distinguishes between "real," meaningful language use and drill activity.

	TEACHER	STUDENT
	TARGET LANGUAGE	
REAL	*A* Evoking student response	*1* Evoking response
	B Evoked by student	*2* Responding
	C Classroom management	
	D Facilitating performance or reinforcing behavior	
	E Information explanation	
DRILL	*G* Evoking stimulus	*3* Individual response
	H Repetition reinforcement	*4* Choral response
	J Prompting	
	P Modeling or correcting	
READING AND WRITING	*W* Presenting written language	*5* Writing
		6 Reading silently
		7 Reading aloud
	ENGLISH	
	K About target structure or sound system	*8* Question about target
		9 Answer about target
	M About meaning	
	N Management	

\+ Silence or English not in the above categories but which seems to facilitate learning
\- Silence or English not in the above categories but which seems to impede learning

At regular intervals of a predetermined number of seconds the observer records a letter or a number, depending on the behavior occurring at that particular instant.

Many types of observational instruments have been developed since Flanders. (See for example Grittner, 1969 and Wragg, 1970). The two described here are mentioned because they have been developed specifically for the foreign language classroom.

Observation instruments have several advantages:

1. They can provide valid and reliable information on certain classroom behaviors when the classroom is observed several times.

2. They are adapted to a variety of tasks, settings, and individuals at all educational levels.

3. They can provide a valuable supplement to achievement data.

4. They can provide both qualitative and quantitative data.

Caution must be exercised, however, whenever one plans on using an observational technique. A long period of training and experience may be required for the observer. In addition, many activities take place simultaneously in a classroom, and it is often difficult to record behaviors that are significant. Interpretation of observational findings must take into account the context and must not generalize from a very restricted sampling of behavior.

REFERENCES

Carroll, J. B. and S. M. Sapon. 1967. Elementary Modern Language Aptitude Test. New York: The Psychological Corporation.

Carroll, J. B. and S. M. Sapon. 1959. Modern Language Aptitude Test (MLAT). New York: The Psychological Corporation.

Common Concepts Foreign Language Test. 1962. Monterey: California Test Bureau.

Cronbach, Lee J. and Patrick Suppes, eds. 1969. Research for tomorrow's schools: Disciplined inquiry for education. Report of the Committee on Educational Research of the National Academy of Education. New York: Macmillan.

Dunkin, Michael J. and Bruce J. Biddle. 1974. The study of teaching. New York: Holt, Rinehart and Winston.

Flanders, N. A. 1960. Interaction analysis in the classroom: A manual for observers. Ann Arbor: University of Michigan.

Grittner, Frank M. 1969. Teaching foreign languages. New York: Harper & Row.

House, Ernest R. 1978. Assumptions underlying evaluation models. Educational Researcher 7 (March): 4-12.

Jakobovits, L. Foreign Language Attitude Questionnaire (FLAQ). 1970. Middlebury, Vermont: Northeast Conference.

Jarvis, Gilbert A. 1968. A behavioral observation system for classroom foreign language skill acquisition activities. The Modern Language Journal 52 (October): 335-41.

Klein, Steven P. 1971. The uses and limitations of standardized tests in meeting the demands of accountability. Los Angeles: University of California, Center for the Study of Evaluation. ED 053 175

Law, Alexander I. and William H. Bronson. 1977. Program evaluator's guide. Sacramento: California State Department of Education; Princeton, New Jersey: Educational Testing Service. ED 142 563

Leinhardt, Gaea. 1977. Program evaluation: An empirical study of individualized instruction. American Educational Research Journal 14 (Summer): 277-93.

MLA Cooperative Foreign Language Tests. 1963. Princeton, New Jersey: Educational Testing Service.

Moskowitz, Gertrude. 1968. The effects of training teachers in interaction analysis. Foreign Language Annals 1 (March): 218-35.

Pimsleur, P. 1966. Pimsleur Language Aptitude Battery. New York: The Psychological Corporation.

_____. 1967. Pimsleur Modern Foreign Language Proficiency Tests. New York: The Psychological Corporation.

Popham, W. James. 1973. Evaluating instruction. Englewood Cliffs, New Jersey: Prentice-Hall.

Scriven, M. 1967. The Methodology of evaluation. In Perspectives of curriculum evaluation, R. E. Stake, ed. AERA Monograph Series on Curriculum Evluation, No. 1. Chicago: Rand McNally.

Stake, Robert E. 1967. The countenance of educational evaluation. Teachers College Record 68: 523-40.

Taylor, Peter A. and Doris M. Cowley, eds. 1972. Readings in curriculum evaluation. Dubuque, Iowa: William C. Brown Company.

TenBrink, Terry D. 1974. Evaluation: A practical guide for teachers. New York: McGraw Hill.

Tuckman, Bruce W. 1979. Evaluating instructional programs. Boston: Allyn and Bacon.

Tyler, Ralph W., Robert M. Gagné and Michael Scriven. 1967. Perspectives of curriculum evaluation. Chicago: Rand McNally.

Worthen, Blaine R. and James R. Sanders. 1973. Educational evaluation: Theory and practice. Worthington, Ohio: C. A. Jones Publishing Company.

Wragg, E. C. 1970. Interaction analysis in the foreign language classroom. The Modern Language Journal 54 (February): 116-20.

Documents identified by an ED number may be read on microfiche at an ERIC library collection or ordered from the ERIC Document Reproduction Service, P.O. Box 190, Arlington, VA 22210.

Gilbert A. Jarvis (Ph.D., Purdue University) is Professor of Foreign Language Education at The Ohio State University. Dr. Jarvis is the editor of Volumes 5, 6, 7, and 8 of the *ACTFL Foreign Language Education Series* and a co-author of six textbooks in French. He contributes regularly to professional journals and has taught French at the elementary, secondary, and college levels.

Shirley Adams (Ph.D. candidate, The Ohio State University) is a Research Associate in Foreign Language Education. She has been involved in the evaluation of various language programs and has taught beginning and intermediate French courses while at Ohio State. Ms. Adams is an associate editor of the *ACTFL Annual Bibliography*.

LANGUAGE IN EDUCATION: THEORY AND PRACTICE

Volume 1 (1977-78)

1. *Directory of Foreign Language Service Organizations*, by Sophia Behrens. $3.95. (ED 153 503)

2. *The Linguist in Speech Pathology*, by Walt Wolfram. $2.95. (ED 153 504)

3. *Graduate Theses and Dissertations in English as a Second Language: 1976-77*, by Stephen Cooper. $2.95. (ED 153 505)

4. *Code Switching and the Classroom Teacher*, by Guadalupe Valdés-Fallis. $2.95. (ED 153 506)

5. *Current Approaches to the Teaching of Grammar in ESL*, by David M. Davidson. $2.95. (ED 154 620)

6. *From the Community to the Classroom: Gathering Second-Language Speech Samples*, by Barbara F. Freed. $2.95. (ED 157 404)

7. *Kinesics and Cross-Cultural Understanding*, by Genelle G. Morain. $2.95. (ED 157 405)

8. *New Perspectives on Teaching Vocabulary*, by Howard H. Keller. $2.95. (ED 157 406)

9. *Teacher Talk: Language in the Classroom*, by Shirley B. Heath. $2.95. (ED 158 575)

10. *Language and Linguistics: Bases for a Curriculum*, by Julia S. Falk. $2.95. (ED 158 576)

11. *Teaching Culture: Strategies and Techniques*, by Robert C. Lafayette. $2.95. (ED 157 407)

12. *Personality and Second Language Learning*, by Virginia D. Hodge. $2.95. (ED 157 408)

Volume 2 (1978-79)

13. *Games and Simulations in the Foreign Language Classroom*, by Alice C. Omaggio. $5.95.

14. *Problems and Teaching Strategies in ESL Composition*, by Ann Raimes. $2.95.

15. *Graduate Theses and Dissertations in English as a Second Language: 1977-78*, by Stephen Cooper. $2.95.

16. *Foreign Languages, English as a Second/Foreign Language, and the U.S. Multinational Corporation*, by Marianne Inman. $4.95.

17. *Testing Oral Communication in the Foreign Language Classroom*, by Walter H. Bartz. $2.95.

18. *Intensive Foreign Language Courses*, by David P. Benseler and Renate A. Schulz. $4.95.

19. *Evaluating a Second Language Program*, by Gilbert A. Jarvis and Shirley J. Adams. $2.95.

20. *Reading a Second Language*, by G. Truett Cates and Janet K. Swaffar. $2.95.

21. *Chicano English*, by Allan A. Metcalf. $2.95.

22. *Adult Vocational ESL*, by Jo Ann Crandall. $5.95.

To subscribe to the complete series of publications, write to the Publications Department, Center for Applied Linguistics, 1611 North Kent Street, Arlington, Virginia 22209. The subscription rate is $32.00 per volume. Titles may also be ordered individually; add $1.75 postage and handling. Virginia residents add 4% sales tax. ALL ORDERS MUST BE PREPAID.